

# The Spread of the 2014 Ebola Zaire Virus in West Africa

Matthew Scotch, PhD, MPH<sup>a</sup>, Rachel Beard<sup>a</sup>, Robert Pahle, PhD<sup>a</sup>, Anuj Mubayi, PhD<sup>a</sup>, Sirish Namilae, PhD<sup>b</sup>, and Ashok Srinivasan, PhD<sup>c</sup>

<sup>a</sup>Arizona State University, Tempe, AZ, USA; <sup>b</sup>Embry-Riddle Aeronautical University, Daytona Beach, FL, USA; <sup>c</sup>Florida State University, Tallahassee, FL, USA

## Abstract

Air travel has been identified as a leading factor in the spread of several infectious diseases. This has motivated calls for limitations on air travel during epidemics, such as during the 2014 Ebola outbreak in West Africa. However, such limitations carry considerable economic and human costs. We are developing a computational framework for public health decision support during infectious disease outbreaks that will enable individuals to make informed real-time decisions that mitigate the risk of spread without eliminating air travel. This multi-component platform models elements of human movement with global diffusion and molecular evolution of the virus. Our human movement model will estimate the number of infected individuals at each airport. We will then use these data in our phylogeography models of global spread to empirically evaluate the impact of air travel policy options such as changes to enplaning, in-flight movement, and deplaning.

Here, we report the initial results of our phylogeography models. We collected full genome 2014-15 Ebola Zaire virus sequences from the Virus Pathogen Resource database (VPR) [1]. We recorded temporal and geographical metadata of each sequence including the decimal day of year and the location of the virus. We aligned 810 sequences using Muscle [2] and selected an HKY +  $\Gamma$  [3] DNA substitution model based on prior phylogeography studies of the 2014 strains [4-6]. We used BEAUti v1.8.2 [7] to specify different assumptions including codon partitions, strict or relaxed molecular clocks, and different tree priors for the Bayesian inference.

We report the preliminary phylogeographic results including the Kullback-Leibler (KL) divergence test [8] for the most likely origin, and the association index for the influence of geography on the genomic evolution of the virus.

## Introduction

In Figure 1, we show the data and information flow for our two models, fine-scale human movement during air travel and global virus phylogeography. The output of our human movement model will be an estimate of the number of infected individuals at each airport. We will then use these data in our phylogeography models of global spread to empirically evaluate the impact of human movement given a series of air travel policy options.

As a first step, we analyzed the phylogeography of the 2014 Ebola outbreak in West Africa. This work extends prior studies [4-6] by considering additional sequences and locations.

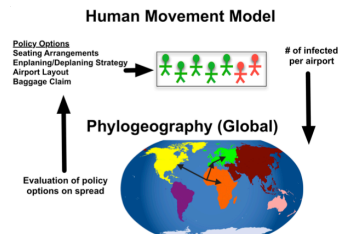


Figure 1. Integration of Models for Ebola Virus Transmission.

## Methods

1. Searched the Virus Pathogen Resource (VPR) by specifying: Species = Ebola Zaire; Minimum genome length = 18,700; Time = 2014-2015; No duplicate sequences = True; Host = Human
2. Removed sequences: Without a collection day, month, and year (N=15); From the Democratic Republic of Congo that were genetically different (N=5); Unverified sequence KP759646
3. Aligned 810 full-genome sequences using Muscle [2]
4. Created different models by molecular clocks, tree priors, and codon partitions
5. Ran Bayesian Markov chain Monte-Carlo simulations 100 million steps in BEAST [7]
6. Evaluated models by computing the log marginal likelihood using Tracer 1.5 [9] Best = HKY+ $\Gamma$  substitution, uncorrelated lognormal relaxed clock, constant tree prior
7. Created a Maximum Clade Credibility tree in TreeAnnotator [7] with a 10% burn-in
8. Created a Google Earth file of the MCC tree using SPREAD [10]
9. Calculated the KL divergence test of the root state using R [11]
10. Calculated the association index of the virus location trait using BaTs [12]
11. Calculated coalescent statistics using PACT [13]

## Results

Figure 2. Maximum Clade Credibility (MCC) Tree.

Guinea had 0.999 posterior probability at the root state (outbreak origin) with a KL of > 5 suggesting strong support. The Time of the Most Recent Common Ancestor (TMRCA) was estimated to be 1.99 years. On the inset, we show the persistence values and 95% credible intervals. This is the average time (in years) for a virus to leave its sampled location [14] and is similar for the African countries.

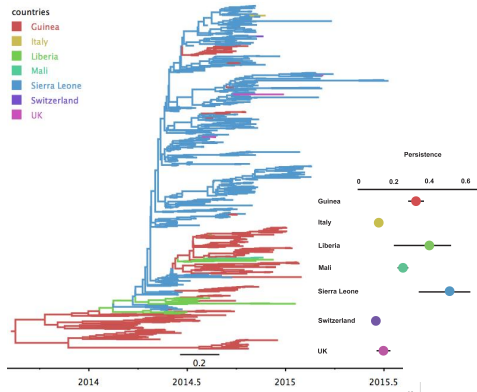


Figure 4. Association Index of the Virus Location Trait.

Bars represent 95% highest posterior density. We reject the null hypothesis that no single tip with a location trait is any more likely to share that location with adjoining taxa than by chance. This suggests that the evolution of Ebola Zaire is structured by geography.

Figure 5. Origin in Guinea.

Here, we show the MCC Tree in Google Earth.



Figure 6. Spread Throughout W. Africa.

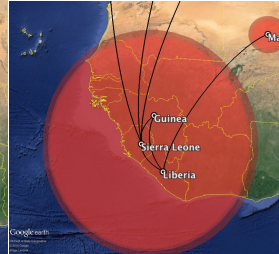


Figure 7. Cases in Europe.

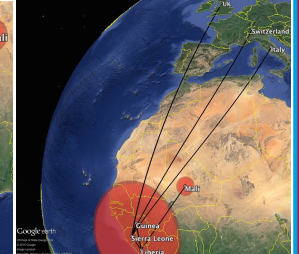
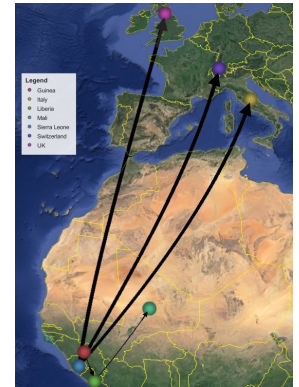


Figure 3. Pairwise Migration Rates between Countries.

The rates are in terms per lineage per year with thicker lines representing greater rates. The high values to Europe (SL → Switzerland is 14.6) are impacted by the small amount of sequences.



## Conclusion

Our findings that Guinea was the origin of the outbreak are consistent with prior results [4-6]. In addition, statistical phylogenetic trait analysis showed that the evolution of the Ebola Zaire viruses was geographically structured while coalescent analysis indicated that there was no difference in virus persistence among the African countries.

Our long term goal is to integrate our global models of virus diffusion with fine-scaled models of human movement on airplanes. In doing this, we will create a computational infrastructure that will study the impact of different policy options aimed at reducing virus spread during air travel. We will adopt a Bayesian phylogeographic generalized linear model (GLM) approach [15] to reconstruct spatiotemporal patterns of viral spread while simultaneously assessing the impact of policy decisions related to air travel such as enplaning and deplaning, seating arrangements, and changes in gathering areas in airports. Our models will enable airlines and transportation agencies to make informed decisions to mitigate the risk of epidemics while minimizing morbidity and mortality of infectious diseases.

## Acknowledgments

Research was supported by the National Science Foundation (NSF) under Award Number ACI 1525012. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF. We thank the National Center for Supercomputing Applications for the use of their Blue Waters supercomputer.



## References

1. Pickett, B. E., et al. (2012). "VPR: an open bioinformatics database and analysis resource for virology research." *Nucleic Acids Res* 40(Database issue): D593-598.
2. Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Res* 32(5): 1792-1797.
3. Hasegawa, M., et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." *J Mol Evol* 22(2): 160-174.
4. Simon-Loriere, E., et al. (2015). "Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic." *Nature* 524(7553): 102-104.
5. Gire, S. K., et al. (2014). "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak." *Science* 345(6202): 1369-1372.
6. Park, D. J., et al. (2015). "Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone." *Cell* 161(7): 1516-1526.
7. Drummond, A. J., et al. (2012). "Bayesian phylogenetics with BEAUti and the BEAST 1.7." *Mol Biol Evol* 29(8): 1969-1973.
8. Kullback, S., Leibler, R.A. (1951). "On information and sufficiency." *Annals of Mathematical Statistics* 22 (1): 79-86.
9. Rambaut, A., et al. (2009). "Tracer." <http://tree.bio.ed.ac.uk/software/tracer/>.
10. Bielejec, F., et al. (2011). "SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics." *Bioinformatics* 27(20): 2910-2912.
11. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
12. Parker, J., et al. (2008). "Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty." *Infect Genet Evol* 8(3): 239-246.
13. Bedford T. (2013). "FACT: Posterior Analysis of Coalescent Trees." <http://bedford.io/projects/FACT/>.
14. Bedford, T., et al. (2015). "Global circulation patterns of seasonal influenza viruses vary with antigenic drift." *Nature* 523(7559): 217-220.
15. Lemey, P., et al. (2014). "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2." *PLoS Pathog* 10(2): e1003932.